North Carolina Digital Repository Requirements

Draft
December 2014

Executive Summary

The following document lays out the requirements of both the State Archives and the State Library of North Carolina for a digital preservation system. The requirements are based on certain use cases specific to the different collections and materials collected, described, and cataloged by the library and the archives. The use cases are listed below and are followed by a numbered list of core requirements categorized as either "required," "preferred," or "optional." These requirements conform to the North Carolina Digital Preservation Policy and are constructed to comply with ISO 14721 (OAIS Reference Model) and ISO 16363 (Trusted Digital Repository Checklist).

Use Cases

State Archives of North Carolina

Born-Digital State Records and Special Collections

Context: State Records Center receives large transfer of born-digital state records.

Goal: An archivist need to appraise, ingest, preserve, arrange, and describe an accession of born-digital records/objects in a way that is efficient and secures the records.

Agents: Processing archivist.

Actions: An archivist brings in an accession of state records on an external hard drive. Information about the accession gathered from the transferring office is input into a records lifecycle management system (in our case, AXAEM), and that system **must** be able to share that information with the digital preservation repository and track the current status of ingested files.

After the accession has been checked for viruses, the archivist creates a manifest of the original transfer. The archivist then culls non-records and non-permanent records before the accession is ingested into the repository. The archivist supplies metadata that describes contextual and technical aspects of the accession, and the repository makes direct use of this metadata as it processes the files. The manifest of the original transfer is retained as part of the final AIP.

The archivist will instruct the repository system to process born-digital objects according to Preservation Workflow A.

The archivist should be able to inspect and, if necessary, correct any metadata harvested by the system and to add metadata, descriptive or otherwise, to the object before it is converted to an AIP. The AIP will consist of the original objects and their original metadata, any added, extracted, or embedded metadata, any normalized derivatives, and file hashes.

Once the AIP is stored, the system will perform regular backups of the files and periodically check them for corruption or obsolescence.

The system must be able to produce a DIP from the contents of the AIP, either at the point of AIP creation or at a later point when the materials are set to be arranged and described.

At any point after the creation of the AIP, the archivist will begin to arrange and describe the materials, using the system to apply descriptive metadata to or associate that metadata with the files in the AIP and their access copies (which make up the DIP).

Preservation Workflow: A. The majority of files in the digital preservation repository will be born-digital. In the case of born-digital state government records and special collections, the repository will provide its full suite of services. (See Appendix, Workflow A.)

North Carolina's Statewide Imagery Program

Context: Center for Geographic Information and Analysis (CGIA) has a deposit of digital geospatial data ready for transfer.

Goal: Archivists ingest a large number of files related to North Carolina's Statewide Imagery Program and assign appropriate preservation actions.

Agents: Digital Services Section staff.

Actions: Originating agency prepares the files for transfer and ensures that the files format complies with the State Archives preservation formats. An archivist receives a transfer of geospatial data (orthophotographs and shapefiles) from NC OneMap Clearinghouse, performs a virus check, and validates the files and the metadata.

An archivist ingests either a large number files at once or in a series of smaller parts that the system reassembles into a single package once the parts have been ingested. Since files are transferred to the archives in preservation formats, the system does not need to create normalized copies.

The system generates access copies from the files in the AIP, unless they are provided by the NC OneMap Clearinghouse. In the latter case, the system will link user-submitted DIPs to system-generated AIPs.

The archivist maps technical metadata extracted by the repository system to descriptions of the objects at the collection, series, file, and/or item level.

Preservation Workflow: B: Current workflows between CGIA and the State Archives take digital preservation into account, providing hi-res TIFFs of all images in addition to geospatial files such as shapefiles (.dbf, .prj, .sbn, .shp, .shx) and MrSID files (.sid, .sdw) that we receive in normalized versions (XML versions of .shp files) or that we consider already to be preservation worthy. Because these files are unique and born digital, they will be replicated and backed up. (See **Appendix, Workflow B**.)

Digitization for Preservation

Context: A patron requests to view or to obtain a digitized copy of a quarter-inch open-reel audio tape that is nearly physically unplayable.

Goal: Archivists need to ensure that digitized copies of damaged or degraded archival or special collections materials remain accessible far into the future since the originals will soon become unusable, if they are not already.

Agents: Audio Visual Materials Archivist.

Actions: A patron requests access to an old open-reel audio tape in one of the archives' special collections. An archivist discovers that the tape is beginning to show signs of degradation. Since the tape would likely not withstand many more (if any) runs through the reel-to-reel, the archivist consults with Digital Services Section and decides that the reel should be digitized in a way that preserves maximum fidelity to the original.

The archivist works with a vendor or with an in-house technician to have the tape digitized to specifications provided by the Digital Services Section—specifications about file format, bit depth and sample rate, and required metadata. This way, the digitized file(s) produced is already "preservation ready" and will require less processing by the repository system.

The archivist ingests the digitized file into the repository, selecting Preservation Policy C, which applies only a subset of the repository's total preservation actions to the file

The archivist will have the option of whether or not to create an access copy, which would depend on the file's copyright status (which in this case is public domain), any donor restrictions, and other factors, which

will be recorded in the PREMIS section of the object's metadata. The system will be able to check that rights metadata automatically or on request for any changes in rights status.

Preservation Workflow: C. Digitization can serve as a means of preservation, and in these cases, the digital surrogate comes to function in place of the original as the object in need of full preservation treatment. (See **Appendix, Workflow C**.)

Large-scale Digitization Projects for Access

Context: Archives staff digitizes a large amount of archival material for wider access online, for an online exhibit, or to commemorate a historical event.

Goal: Archivists need to be able to apply a minimal or light-weight version of the repository's preservation policies to digitized records/objects that do not warrant full format preservation.

Agents: Archivists in the Digital Access Branch of the Digital Services Section.

Actions: A large volume of archival materials has been selected for mass digitization for online access and display due to its historic importance and wide cultural appeal. The materials are in fair-to-good shape, so they are in no immediate danger of becoming inaccessible. However, since the digital files are of considerable interest to the public and taking into account the amount of resources spent to digitize the original records, these digitized surrogates represent an investment of significant public and cultural value and require a higher level of preservation than that usually afforded to access copies generated upon patron request. Yet since the digital files are not unique (the originals still exist and are useable), the full range of preservation activities is unnecessary.

The originals are digitized to preservation standards, so no normalization needs to take place, and the archivist ingests these files under Preservation Policy D. The archivist has also created metadata that need to be associated with the files. The archivist has also created access copies (JPEGs) that need to be ingested and recognized by the repository system as such.

Preservation Workflow: D. Some digitization projects are not motivated by preservation concerns and yet constitute a significant resource and investment for the library and archives. The authenticity of the files is less of an issue than with born-digital files and there may be no need for replication, incremental backups may be sufficient. (See **Appendix, Workflow D**.)

State Library of North Carolina

Deposit Born-Digital State Publications that have not been cataloged, via CINCH

Context: A library technician has a CINCH zip package of online state agency publications.

Goal: Identify and ingest uncataloged publications, then pass them to the library's catalogers for cataloging and DIP dissemination.

Agents: Library technician

Actions: The library technician submits the CINCH zip package to the system as a pre-ingest package. The system needs to use the CINCH-supplied metadata as a basis for the files' metadata records. It will perform automated ingest actions on the publication files according to **Preservation Workflow B**.

Next, the technician will need to manually process and describe the files. First, she will determine whether each file is a duplicate of a publication already received by the Library, which she will determine based on (1) checksum de-duping and (2) manual comparison of titles and publication date. Next, she will review whether to discard additional publications.

The system should facilitate the technician's ability to easily identify duplicate publications, discard duplicates, and discard other publications. There is minimal need to keep a record of those files that have been discarded.

For remaining publications, the technician will need to identify whether each file is a stand-alone item (a monograph) or part of a serial. She will then manually identify the publication date and/or issue number. The system will need to provide her a way to associate the file with a serial.

Next, she will need to determine if the serial has been cataloged. For those that have, see the following use case. For those that have not been cataloged, she needs the system to alert the cataloging staff that the new serial is ready to be cataloged and made publicly available.

Preservation Workflow: B. (See Appendix, Workflow B.)

Catalog born-digital State Publications

Context: A library technician has ingested born-digital state publications and identified one or more as unique (uncataloged) publications. An alert has been received by the cataloging librarians.

Goal: Assign publications to a cataloger, who catalogs the publications, generates a DIP, and exposes the DIP through the access system (CONTENTdm)

Agents: Catalogers

Actions: After the technician has ingested digital issues of an uncataloged serial or monograph, the system will alert cataloging staff that the files are ready to be cataloged and uploaded to the access system (CONTENTdm).

A cataloger will "check out" each publication. Then, she will use OCLC's Connexion Client to catalog the publication in the Library's local catalog and in WorldCat. She will generate a DIP using metadata generated during the ingest process, metadata generated in the Connexion Client, and possibly additional metadata generated by the cataloger. The cataloger may edit metadata generated during the ingest process. She will then upload the DIP to CONTENTdm.

Note: in nearly all cases, the preservation and access copies are the same PDF (i.e., the AIP Content Data Object: = DIP Content Data Object), thus the system need not necessarily store a separate access copy of the publication PDF.

The DIP metadata will need to be formatted according to the access system's specifications (CONTENTdm). The DIP is uploaded to CONTENTdm as a new object. The digital repository system must be able to integrate with this workflow such that:

1. The system stores or has access to the canonical metadata for each digital publication, including all metadata pushed to the catalogs and access system (CONTENTdm) and any additional preservation and administrative metadata.

2. The system can report the status of any given publication and whether it has been cataloged and published to the access system (CONTENTdm)

Preservation Workflow: n/a

Publish born-digital State Publication Serials that have already been cataloged

Context: A library technician has processed an ingest and identified some files as belonging to already-cataloged serials.

Goal: Make publicly available any publications that have already been cataloged. Publications that have already been cataloged include new issues of pre-existing serials.

Agents: Library technician

Actions: The technician will generate a DIP using the metadata generated during the ingest process and additional metadata from the existing catalog record.

Note that in nearly all cases, the preservation and access copies are the same PDF (i.e., the AIP Content Data Object: = DIP Content Data Object), thus the system need not necessarily store a separate access copy of the publication PDF.

The DIP metadata will need to be formatted according to the access system's specifications (CONTENTdm). The DIP is uploaded to CONTENTdm, with new issues being added to the serial's existing CONTENTdm compound object.

The system then needs to add a new date to the publications existing catalog entry, and publish it to the Library's catalog (Voyager) and WorldCat upon approval from a cataloger. Optionally, the system could alert the cataloging staff that a new date range should be added to the publications' existing catalog entry. Either way, the digital repository system must be able to integrate with this workflow such that:

1. The system stores or has access to the canonical metadata for each digital publication, including all metadata pushed to the catalogs and access system (CONTENTdm) and any additional preservation and administrative metadata.
2. The system can report the status of any given publication and whether it has been cataloged and published to the access system (CONTENTdm)/

Preservation Workflow: n/a

Deposit Digitized Surrogates of SLNC-owned Materials

Context: A vendor has digitized a shipment of the Library's books and has made the PDFs available for download via Internet Archive. A SLNC tool facilitates one-click downloading of the files from the Internet Archive.

Goal: Retrieve the PDF files, ingest them into the digital repository system, edit the associated metadata, and make the publications publicly available in CONTENTdm.

Agents: Digital projects librarian

Actions: A set of PDF files and Internet Archive–formatted CSV metadata has been made available for download, and the digital projects librarian has been alerted to its availability.

She needs to ingest the Internet Archive package into the digital repository system, where she will spot check the quality of the PDF publications, associate the publications with an existing catalog record, add additional metadata, generate DIPs, and upload the DIPs to the access system (CONTENTdm).

Note that in nearly all cases, the preservation and access copies are the same PDF (i.e., the AIP Content Data Object: = DIP Content Data Object), thus the system need not necessarily store a separate access copy of the publication PDF.

Preservation Workflow: B. (See Appendix, Workflow B.)

Shared Requirements

Fixity check of digital assets

Context: Regularly-scheduled bit-level preservation of all digital assets

Goal: Identify any bit level errors/changes in AIPs, and repair with a copy from another storage copy.

Agents: Electronic Records Archivist, Systems Librarian, Digital Collections Manager

Actions: The system runs checksums across all AIP data content objects in the repository's local preservation storage twice each year, comparing the current hash value of each file to the file's original token. Where files are found to have matching hash values, a PREMIS event should be recorded, and the redundant copy of the hash value need not be stored.

Where files are found to have a conflicting hash value, the system checks backup storage, confirms that a copy is present in backup storage and that its hash matches the original hash value, and replaces the local copy with the backup copy.

Stringent protocols **must** be in place to ensure that intact local copies of objects are not inadvertently overwritten by compromised copies from backup storage, including possibly:

1. hashing of checksums themselves, to ensure that the original hash value has not been corrupted or tampered with and
2. protocols that require administrator approval before any local files are overwritten from backup.

AIP metadata (or, "Preservation Description Information") requires similar auditing and repair procedures, although the nature of these procedures will depend on how the system manages the metadata (e.g., as decentralized XML files or in a centralized database).

Perform a preservation action on a subset of items

Context: The State Library has determined that it JPEG2000 fmt/151 is a more appropriate preservation format than TIFF for its collection of TIFF book scans. It has also decided to wait 1 year after generating the JPEG2000 files before deleting its TIFF files.

Goal: Identify the computing resources required to migrate all of the State Library's AIPs TIFF objects that represent digitized books, into the JPEG2000 fmt/151 format; migrate the files; and retain the original TIFFs alongside the JPEG2000 for a period of 1 year.

Agents: Digital projects librarian, Digital collections manager

Actions: State Library staff query the system to determine the number and size of all TIFF AIP objects that represent digitized books, and they tell the system that they would like to perform a test run to migrate the identified files into the JPEG2000 format. The system outputs the intended log of actions and results. State Library and State Archives staff coordinate the timing of other processes, and schedule a time in the system for the State Library's targeted files to be migrated. Staff indicates that the TIFF originals should not be deleted for another year.

After migration is complete, the new JPEG2000 files are associated alongside their TIFF counterparts within the same AIP. Any errors are reported to staff, who may then attempt to manually convert files and associate them with the intended AIP.

The TIFF files are also uploaded to backup storage.

In one year's time, system administrators are alerted that it is time to delete the original TIFFs, and they are prompted to initiate the deletions. The TIFFs are also removed from backup storage.

Determine whether a given title is in preservation storage and has been added to the access portal (CONTENTdm)

Context: A reference librarian has received a request from a state agency employee who remembers sending the Fall 2013 issue of "Cat's Pause" to the State Library, but can't find the issue in the access system (CONTENTdm). The reference librarian also cannot find the issue, and emails the digital collections manager to check.

Goal: Identify whether the publication has been received, what its processing status is, and if there is a reason it is not publicly available.

Agents: Digital collections manager

Actions: The digital collection manager has the title and issue date of the publication in hand ("Cat's Pause," Fall 2013) and is also unable to find it in the access system (CONTENTdm). She queries the digital repository system using the publication title and issue date, and finds that the issue has been received via CINCH but that it has not been fully processed by the library technician and thus has not yet been uploaded to CONTENTdm.

Core Requirements

Security

Level 1: Required

1. Repository system allows for assigning of user roles with varying levels of security access.
2. Repository system records all user actions in the system for auditing.
3. Repository system has sufficient security protocols to protect storage from malicious attack.

Pre-Ingest

Pre-ingest includes the processes by which the system recognizes files and compiles them into ingestible SIPs

Level 1: Required

1. Staff deposit using BagIt: Repository system validates a submitted bag using the bag's manifest file
2. Staff deposit of unstructured files: Repository system can take files without any metadata, create a checksum hash of file in place, copy file into system storage, and validate checksum
3. Repository system performs virus checks on all files prior to ingest and maintain an up-to-date virus dictionary
4. Staff can reject a pre-ingest package, or portion of pre-ingest package

Level 2: Preferred

1. Creator deposit: Creators (state agencies, local agencies) can deposit objects, with accompanying metadata, through web forms for review and eventual ingest into system by archival staff
2. Appraisal: Archivists are able to review and cull groups of records before ingest using
3. a tool that maintains the integrity of the files and secures them from accidental deletion

Level 3: Optional

1. Staff deposit using CINCH. Repository system validates file checksums in CINCH metadata file, checks for CINCH event, error and file metadata, and allows system operator to make decisions about CINCH-identified "problem files" based on error messages in CINCH metadata
2. Creator deposit, CINCH URLs: Creators (state agencies) can deposit URLs to state agency publications

Ingest

Level 1: Required

1. Repository system maintains SIP content data within AIP unless instructed otherwise by archivist
2. Repository system performs uniqueness check based on checksum hash: System checks current deposits against all previous deposits' checksums. In the event of a collision, the item is flagged for review
3. Repository system ingests every new SIP or file as unique (allows for duplicates with no collisions)

4. Repository system allows for separate ingest streams with different levels of preservation actions
5. Repository system communicates with the library and archives' data management system AXAEM, linking files and information within the repository to accession, deposit, gift, and donor information in the data management system
6. Repository system has configurable rules for handling virus-positive data streams (e.g., quarantine, delete/destroy, alerts, document in PREMIS metadata)
7. Repository system validates fixity values delivered via BagIt (SHA256)
8. Repository system validates fixity values delivered via CINCH CSV files (SHA1)
9. Repository system creates fixity values for objects deposited without fixity values
10. Repository system stores all fixity values are apart from AIP, so that accidental alteration of the AIP would not also damage the fixity information
11. Repository system inventories original file tree & file names
12. Repository system normalizes/cleans file names
13. Repository system extracts zip and other compressed packages
14. Repository system characterizes file formats and confirms files are valid and well formed (error rate should be in line with industry standard tools).
15. Repository system extracts embedded metadata
16. Repository system generates basic metadata (file size, deposit date, file name)
17. Repository system normalizes file formats according to admin-configurable rules
18. Repository system clears file system permissions (read-write-execute/access control lists)
19. Repository system identifies digital locks & handle according to admin-configurable rules (e.g., alerts, remove locks)
20. Repository system can mitigate Window's limit on file structure depth/length of file paths.
21. Repository system allows for creation of basic provenance and rights metadata (e.g., indicating receipt of paperwork on custody, accession, ownership, copyright, etc.)
22. Repository system allows staff to spot check AIP contents to ensure that files are understandable, i.e., can be accessed, opened, and rendered for human understandability

Level 2: Preferred

1. Ingest produces a report that can be returned to the donor/transferring agency that indicates what files were ingested and which were ingested successfully and which were not successfully ingested, or that had invalid checksums
2. If SIP processing transformations occur in a secondary tool outside of the repository system, the system is able to receive the metadata produced by the secondary tool to document transformations applied to the SIP during its construction into an AIP

Intellectual Control/Cataloging

Level 1: Required

1. Users can query system to locate assets based on file and AIP level metadata.
2. Repository system generates and continually records PREMIS metadata associated with each file/collection of files and makes that data available through reports or web interface
3. Repository system allows for the creation and updating of descriptive metadata (MODS) for archival holdings at any level of granularity—collection, group, series, and file—allowing for an archivist to perform arrangement and description long after AIPs have been created and preserved
4. Repository system must share descriptive metadata with access system in a way that reflects any archival arrangement encoded in that metadata
5. Repository system records original file names as metadata associated with renamed records/objects
6. Repository system stores/links the canonical metadata for each digital asset, at minimum linking to all metadata published to the catalogs (AXAEM and Voyager) and storing all metadata published to the access system (CONTENTdm), including any additional preservation and administrative metadata
7. All metadata schemas used by the repository system are open and documented

Level 2: Preferred

1. Repository system should allow the Archives and Library to define a minimum set of descriptive metadata that must accompany every AIP for the AIP to be valid and complete.

Active Preservation/Long Term management

Level 1: Required

2. Repository system performs system-wide, scheduled fixity checks with logs/reports at least twice per year, where AIP data objects and metadata are checked and synced all storage providers (cloud and otherwise)
3. Hash value tables/file receives regular fixity checks
4. Repository system gives clear indication of files affected by any bit-level changes
5. Repository system can correct or recover altered or corrupt files, with stringent protocols in effect to prevent accidental replacement of intact files with bit-altered files
6. Where unrecoverable bit errors occur in AIPs, repository system permanently retains documentation of those errors such that reports listing all errors can be produced at any time by system administrators
7. Repository system creates multiple redundant backups of repository data, at least one of which is maintained on North Carolina servers. (Repository system may either provide backup storage or be able to integrate with third-party storage)
8. Repository system monitors and migrates file formats
9. Repository system allows for changes to AIPs and tracks all changes to AIPs
10. Repository system migrates file formats based on PRONOM or other file format registries

11. Repository system and developer demonstrate reasonable grounds to expect sustainability of system software over the medium term and ability for infrastructure to evolve gracefully over time in response to evolving technologies

Level 2: Preferred

1. Repository system can query system for detailed reports of repository content, including file size, checksum hash, location, and number of copies, etc. of all materials in the digital repository
2. Repository system allows for AIP versioning, such that migrated file format versions may be associated with the same AIP
3. After file format migration, repository system logs and reports any migration errors

Level 3: Optional

1. Repository system validates that its metadata is synced with external access system metadata (i.e., CONTENTdm)

Access

Level 1: Required

2. Repository system exports DIPs along with associated metadata to the access system
3. Repository system communicates with the library and archives' data management system AXAEM
4. Repository system can embargo files
5. Repository system allows for redaction of DIP or for the upload of a redacted file
6. Repository system can restrict access to files based on metadata (e.g., PREMIS rights metadata)
7. If errors occur during AIP to DIP production, repository system notifies user
8. DIPs can be reprocessed on demand.

Level 2: Preferred

1. Repository system exports DIPs and associated metadata according to specific formats required by access system

Level 3: Optional

1. Repository system allows for one-click publication of DIPs and associated metadata to access system

Exit Strategy

Level 1: Required

2. The repository system contains a clear definition of how AIP components—i.e., content information (the primary data/records being preserved); fixity values; provenance information; reference information (which allows objects to be findable and might include such descriptive information as title and subject terms); access rights information such as confidentiality, type of confidentiality, copyright, other intellectual property rights, embargos; file format information, including format version, validity, well-formedness—are associated, so that the component parts can be identified, associated with the AIP, and parsed from the AIP without the system in place. The AIP components

may be stored in a file system, database, or combination of the two—so long as the information structures are defined for identification and parsing

Appendix

Preservation Workflow A

3. Virus scan
4. Assign persistent identifier
5. Checksum files with SHA-256
6. Recognize and record file formats
7. Validate file formats
8. Extract technical metadata
9. Record when formats cannot be recognized and when formats are incomplete
10. Perform bit-level preservation
11. Create normalized copy based on preservation policies informed by file format registries such as PRONOM and the Unified Digital Format Registry
12. Send AIP to the local storage partition that is replicated at the Western Records Office and backed up to geographically distributed network of servers.
13. Perform regular integrity checks. (Bi-Annual)
14. Maintain PREMIS records of all audits and any changes made either intentionally or by accident to the AIP.
15. Monitor changes to file format (PRONOM and UDFR)

Preservation Workflow B

1. All of Workflow A actions, except that files are not normalized.

Preservation Workflow C

1. All of Workflow A actions, except that SHA-256 is not required (SHA-1 or MD5 would be sufficient), checksums are less frequent, and files are not normalized.

Preservation Workflow D

1. All of Policy A actions are performed for master TIFFs, except that SHA-256 is not required (SHA-1 or MD5 would be sufficient), checksums are less frequent, files are not normalized, and files are backed up but not replicated on Western Records Office servers.